BOYKIN, RONALD AUBREY. The Effects of Instructions and Calculation Procedures on the Accuracy, Agreement, and Calculation Correctness of Observers. (1977)
Directed by: Dr. Rosemery O. Nelson. Pp. 91.

Behavioral research and therapy often rely on data collected in natural settings by human observers. Research interests have begun to focus on the methodological issues involved when data are collected in this manner. These methodological issues have been categorized into four areas: code complexity; observee reactivity; the measurement of inter-observer agreement; and observer bias.

The present study looked at issues related to measuring inter-observer agreement, and attempted to answer the following questions: first, what effect do instructions to observers have on the levels of agreement and observational accuracy they achieve; and second, do instructions observers follow in calculating agreement result in their calculating their own agreement differently than they calculate the agreement levels of other observers.

Sixteen undergraduates were trained to use a behavioral code to record from videotapes the classroom behavior of two eight-year-old second grade males, and to calculate inter-observer agreement levels. The subjects were then paired and randomly assigned to one of the two instructions groups. Instructions were to try to reach agreement of .85 or better with one's partner, or to make one's observational recordings as carefully as possible.

During each of eight experimental trials, observers recorded the occurrence of three target behaviors of one of the children from a 10-minute segment of videotape. Each member of the pair then calculated agreement figures for the target behaviors on two different sets of data: their own data for that trial, and data they were led to believe were collected by another pair of observers in the study.

Four dependent measures were examined: inter-observer agreement; observer accuracy, obtained by comparing each observer's record to a protocol representing the "true" occurrence of the behaviors; the difference score derived by subtracting accuracy from agreement; and a measure of calculation errors derived by subtracting experimenter-calculated agreement scores from observer-calculated agreement scores.

One major finding of the study was that the observers, as a single group, made calculation errors which spuriously inflated their own reported agreement levels, and which spuriously deflated the agreement levels they reported for "other" observers.

The other major finding was that instructions significantly affected the difference between inter-observer agreement and observer accuracy. The group instructed to try to achieve .85 agreement levels produced higher agreement than accuracy; for the group instructed to make their observational recordings and calculations carefully, accuracy was higher than agreement.

The implications of the findings and suggestions for improving research methodology are discussed in Chapter IV.

THE EFFECTS OF INSTRUCTIONS AND CALCULATION PROCEDURES ON THE ACCURACY, AGREEMENT, AND CALCULATION CORRECTNESS OF OBSERVERS

by

Ronald A. Boykin

A Thesis Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
1977

Approved by

Rosemary O. Nelson
Thesis Adviser

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of the Graduate School at the University of North Carolina at Greensboro.

Thesis Adviser Rosemary O. Nelson

Committee Members Jacquelyn Sachelliu

12-20-77

Date of Acceptance by Committee

Acknowledgments

I would like to thank the following people for their contributions to this work: Dr. Rosemery Nelson, who assisted me in every area of the study and who has greatly influenced my thinking and my work over the last two years; Dr. Jackie Gaebelein, who was especially helpful in the design of the study and the analysis of the data; Dr. John Seta, who helped stimulate my thinking in this area of research, and contributed to the interpretation of the results; Skip Beck, who helped me formulate the original research question; the students who served as subjects in the study, for their hard work; to Ms. Elizabeth Hunt, who typed the manuscript; and to my wife Terri, whose encouragement and patience can only be partially repaid by my dedicating this work to her.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

The emphasis on observable behavior is a primary tenet of behaviorism. Behavioral research relies on observational data to study functional relationships between environmental events and behavior. Behavior therapy relies on observational recordings to assess human problems and to evaluate the effects of various types of treatment on remediating these problems.

> If a behaviorist wants to convince someone of the correctness of his approach to treating human problems, he is generally much less likely to rely on logic, authority, or personal testimonials to persuade than are proponents of other schools of psychotherapeutic thought. Rather, it is most likely that he will show his behavioral data with the intimation that this data speaks eloquently for itself. (Johnson & Bolstad, 1973, p. 7)

The primary advantage of using behavioral data is that the data are objective and non-inferential. Unlike traditional psychotherapists, behavior therapists infer no underlying cause of the observed behavior. Also, because they are generally aware that behavior is unique and specific to the setting in which it occurs (Mischel, 1968), behavior therapists seldom infer that behavior observed in one setting will necessarily be observed in any other setting. Finally, because they are generally aware that a person's verbal report of his or her behavior might not be congruent with an

observational record of that behavior, behavior therapists will usually collect data, when feasible, on the behavior as it is occurring in the person's natural environment.

Potential Problems in Using Data Collected by Human Observers

Research utilizing observational recording as a data base has greatly exceeded research on the variables that affect the quality of such data. However, available research indicates that a number of factors may create problems affecting the quality of observational data. It is important to study the nature of these problems and the potential effects they have on the quality of observational data. The internal validity of an experiment may be affected if changes in the dependent variable cannot be attributed solely to the independent variable under study; that is, if the observational procedure itself also contributes to changes in the dependent variable. On the other hand, even under circumstances when no confound exists to affect the internal validity of the results, if the observational procedure is an active ingredient, (i.e., if the observational procedure itself contributes to the variance in scores on the dependent variable) the experimental results may not be generalizable to other settings when no observers are used.

The factors creating problems that affect observational data may be divided into four basic areas: code characteristics; reactivity; procedures used to calculate inter-observer agreement; and observer bias.

Regarding code characteristics, a number of behavior coding systems are available which allow an observer to observe and record the occurrence of several behaviors simultaneously. Mash and McElwee (1974) found that the complexity of such coding systems has an effect on the accuracy of observers. The authors trained observers to use one of two coding systems: the first, a four-behavior coding system; the second, an eight-behavior system, devised by subdividing each of the four categories of the first code into two categories. Observers using the four-behavior coding system achieved significantly higher accuracy than observers using the eight-behavior coding system.

Another issue related to code characteristics is the degree to which the coding procedure permits observers to record behavioral events which are representative of actual ongoing behavior. For example, Thomson, Holmberg, and Baer (1974) compared three different methods of intermittent time-sampling (continuous, alternating, and sequential) of the behavior of four subjects (two teachers and two students) to a continuous time-sampling procedure. The authors found that the intermittent procedure which used the smallest time intervals and progressed sequentially among all four subjects produced data most similar to those produced by the continuous or ongoing time-sampling procedure.

The second major problem that may affect the quality of observational data relates to the effects of the observation

procedure itself on the behavior of those being observed. This phenomenon, usually referred to as "reactivity," may present problems during assessment by changing the frequency of behavior in much the same way as treatment might. Researchers have studied this problem in two ways: either a between-subjects or a within-subjects design may be used to compare the behavior of subjects aware that their behavior was being recorded in a certain setting to the same behavior of subjects unaware that they were being observed and their behavior recorded. Using a between-subjects design, Bechtel (1967) studied time spent and movement around a museum room. People aware that these aspects of their behavior were being recorded spent less time in and made less movement around the room than did people unaware that their behavior was being recorded. Using a within-subjects design, Roberts and Renzaglia (1965) studied subjects' self-comments, first covertly, then after informing the subjects that their verbalizations were being recorded. Subjects made more favorable comments and fewer unfavorable comments about themselves in the overt (informed) condition than in the covert condition.

The third category of problems that may affect observational data relates to procedures for assessing interobserver agreement (also called reliability), or "agreement between observers who independently score the same behavior of a subject" (Kazdin, 1977, p. 141). "A demonstration of high reliability is critical to conclude that a strong relationship

exists between the behavior emitted by the subject and the behavior recorded by the observer" (Lipinski & Nelson, 1974, p. 343). There is a dependency on inter-observer agreement to evaluate the quality of observational data because in most cases no record of the actual behavior of a subject exists. Observing procedures, as well as the conditions under which inter-observer agreement is calculated, may affect the level of agreement obtained, and thus the validity of the observational data themselves. This issue will be elaborated later in this introduction.

The fourth category of problems is related to the potential effects of observer biases in the observational setting which may result in biased or unrepresentative data. Observer bias has been cited as the most pervasive methodological problem in behavioral research (Pawlicki, 1970). Unlike reactive effects which appear either to dissipate over time or to be constant across experimental conditions, the effects of observer bias may interact with treatment conditions to confound results (Rosenthal, 1966), or yield results which cannot be clearly attributed to the independent variable. Since bias threatens the internal validity of experiments, it is important to understand the conditions that produce or minimize observer bias.

Early Research on Experimenter Bias Effects

Rosenthal's (1966) conceptualization of experimenter bias has provided the framework for much of the recent work

in this area. Bias is said to occur when "experimenter effect or error is assymmetrically distributed about the 'correct' or 'true' value" (Johnson & Bolstad, 1973, p. 12). The error that contributes to a score or value of the dependent variable is thought to be random error; thus, bias is assumed to have occurred when observational errors are consistently found in one direction only.

Early research indicated a relationship between biased observational records and a category of antecedent events typically referred to as "expectation biases." Two of Rosenthal's studies are illustrative.

In the first study, Rosenthal and Fode (1963) randomly assigned naive rats to two groups of undergraduate experimenters. One group was informed that their rats were "maze-bright"; the other group was told their rats were "maze-dull." The experimenters were asked to record their rats' running times through a maze. The group of experimenters led to believe their rats were "maze-bright" reported significantly faster running times for their animals than those experimenters led to believe their animals were "maze-dull."

Rosenthal and Jacobsen (1964) randomly selected two groups of classroom students and informed teachers that children in one group were "late-bloomers" (could be expected to show academic gains later in the school year). Fall and spring intelligence testing revealed that those children labelled "late-bloomers" had made greater gains in IQ scores than had children in the control group.

Rosenthal's demonstrations of the effects of experimenters' biases generated much interest in the scientific community. Barber and Silver (1968) reviewed much of Rosenthal's research carefully, and suggested that methodological and statistical problems, as well as the inability or failure of researchers to replicate his results, made some of his findings questionable. The authors concluded that evidence for experimenter bias effects was not as conclusive as was first thought. Barber (1976) suggested that the results of the Rosenthal and Fode (1963) and Rosenthal and Jacobsen (1964) studies may have been due, not to experimenter expectancy effects, but to problems he labels and describes as an "Experimenter Failure to Follow the Procedure Effect, an Experimenter Misrecording Effect, or an Experimenter Fudging Effect" (p. 68). The controversy that followed resulted in a new interest in this area of research, both in the social psychology and behavioral areas.

Behavioral Research on Observer Bias

Several studies have investigated the extent to which informing the observers of the experimental hypothesis or predicted results may result in biased observational records. Scott, Burton, and Yarrow (1967) reported the results of a treatment study in which Dr. Scott, who was informed of the study's predicted results, compared her observational record with those of two uninformed observers. All the observers recorded children's behavior from audio tapes as "positive"

or "negative." The observations of the informed observer differed significantly from those of the uninformed observers, and were in the direction of the experimental hypothesis. Interpretation of these results is limited, however, by the fact that Dr. Scott may also have differed from the uninformed observers in other ways which might have produced the observed discrepancies.

Kass and O'Leary (Note 1) studied, in a simulated field experimental situation, the effects of informing trained observers of expected results on the observational records of these observers. Three groups of observers were trained to use the O'Leary Classroom Behavior Code (O'Leary, Romanczyk, Kass, Dietz, & Santagrossi, Note 2), a code which allows observers to record the occurrence of nine disruptive behaviors of children in the classroom. The observers were then told they would view on videotape the baseline and treatment phases of a classroom intervention program (teacher reprimands). Those in Group One were told that teacher reprimands would result in an increase in disruptive behavior of the target children; observers in Group Two were told to expect a decrease in disruptive behavior; and Group Three observers were given no specific expectation regarding any change in the children's behavior. Protocols or standards for the tapes, previously developed by having trained observers code behavior from the tapes several times until a consensus of agreement was achieved, revealed that the children's disruptive behavior decreased slightly from baseline to treatment.

Observers recorded behavior during four baseline and five treatment sessions. The results generally revealed that those in Group One (told to expect an increase) produced data which showed smaller decreases in disruptive behavior across treatment sessions than those in Group Two (told to expect a decrease) or Group Three (given no expectation). However, the study contained a number of methodological problems: most notably, observer groups were trained separately, and did not compute inter-observer agreement between groups. This could mean that groups of observers may have developed their own interpretation of the behavior code. Johnson and Bolstad (1973) refer to this process as "observer drift." Since expectation conditions were confounded with specific observer groups, the results obtained by Kass and O'Leary may have been due to observer drift rather than differential expectations.

Skindrud (Note 3) and Kent (Note 4) both attempted to replicate the findings of Kass and O'Leary (Note 1) while at the same time improving the methodology of that earlier study. Skindrud trained observers together, then divided them into three groups before having them code videotapes of family interactions. Those in Group One were informed that the target child's behavior would be more deviant when the father was absent than when he was present. Observers in Group Two were told to expect just the opposite (less deviant behavior by the child in the absence of the father

than in his presence). Group Three observers were given no specific expectation. The observational recordings of the three groups revealed no significant differences; moreover, a comparison of each group's recordings with a criterion protocol revealed that the accuracy of recording by each group was equivalent.

Kent (Note 4) trained forty observers as a single group before assigning them to one of eight expectation conditions. The observers were then trained for three more days within their different expectancy groups. Within-group agreement levels of .70 were reached (compared to the .60 average agreement level obtained for the entire group during training); however, again each group seemed to have drifted in its interpretation of the behavioral code prior to the experimental manipulation. Although the results of the study were that the groups' records were not influenced by the expectancies given to them, this finding is open to question because of possible observer drift.

Kent, O'Leary, Diament, and Dietz (1974) trained five pairs of observers to code the occurrence of the nine behaviors of the O'Leary code from videotapes. They then assigned each member of the pair to one of two expectancy groups. Those in Group One were informed that a decrease in disruptive behavior would occur from baseline to treatment, while observers in Group Two were told to expect no change in the children's disruptive behavior from baseline to treatment.

Following the observational recording sessions, observers were asked to give their impressions of what, if any, change in the children's behavior took place from baseline to treatment. The results of the study revealed an interesting contrast. The behavioral recordings of the two groups did not differ significantly; however, the subjective impressions of the two groups differed dramatically: nine of ten observers in Group One (expect a decrease in disruptive behavior) reported they thought the rate of disruptive behavior had decreased; seven of ten Group Two (expect no change) observers reported that they felt no change in disruptive behavior had occurred. Thus, while the quantitative behavioral recordings were not influenced by expectancy, the overall qualitative ratings were influenced to a great degree.

Similar results have been obtained in a study by Shuller and McNamara (1976) in which groups of observers were given different trait labels: normal, hyperactive, aggressive, and no label or expectancy, for the same child whose behavior they were to record from videotapes. Following the observation sessions, observers filled out a rating scale designed to assess their subjective impression of the target child's behavior. The rating scale included dimensions which were related to each of the trait labels given to observers. The global ratings of observers in each group reflected the expectancy they had been given. That is, observers told the child was hyperactive rated the child higher on those dimensions

considered related to hyperactivity than did those in any of the other groups. The behavioral recordings, based on an eight-behavior code which included behaviors the authors considered to be reflective of each of the trait labels provided to the observers, revealed no significant differences among the four expectancy groups on any of the behaviors.

The Kent et al. (1974) and Shuller and McNamara (1976) studies, then, suggest that the effects of observer bias may depend on the response mode studied. In both studies, antecedent conditions (expectancies and trait labels, respectively) failed to bias the observer's quantitative behavioral recordings; on the other hand, both studies revealed that the subjective verbal report of observers could be biased by expectancies. Fogel (Note 5) has suggested that, because in these studies the two types of response mode (verbal report and overt behavioral recording) may not be independent of each other, it would be interesting to see if observers who were given an expectancy and who observed the child in the experimental setting, but did not record behavior, would evidence bias in their verbal report to the same extent as observers who actually record the behavior of the child. She has planned a study to try to answer this question.

While antecedent conditions generally failed to bias behavioral recordings, consequent conditions have been demonstrated to result in biased recordings. O'Leary, Kent, and Kanowitz (1975) have demonstrated the effects on observers'

data of the experimenter's providing evaluative feedback to observers regarding the conformity between their behavioral records and the experimenter's predictions. The authors trained four observers to code from videotapes four behaviors of children in a classroom. The observers were then told that they would view the baseline and treatment phases of a behavior management program, and that a specified two of the behaviors would show a dramatic decrease in frequency from baseline to treatment. The tapes in fact showed no change in any of the four behaviors from baseline to treatment. Following each recording session the experimenter discussed each observer's record with her; during the treatment phase an observer was given positive feedback if her record showed a decrease in the two specified behaviors, and negative feedback if her record of these behaviors showed either no change or an increase from baseline to treatment. This manipulation resulted in the observers' recording a 38% decrease in one specified behavior, a 27% decrease in the other, and essentially no change in the frequencies of either of the control behaviors from baseline to treatment. The results of this study suggest that consequent conditions may bias the behavioral recordings of observers.

Variables Affecting Levels of Inter-observer Agreement Obtained by Human Observers

The research regarding the effects of observer bias on the collection of observational data has been examined above. A problem noted earlier relates to variables that affect the

level of inter-observer agreement obtained by observer pairs. Inter-observer agreement (sometimes referred to as observer reliability) refers to the degree to which two observers of the same behavior concur in their recording of that behavior. Inter-observer agreement may be calculated in several ways, depending on the topography of the behavior and the observational method used to record the behavior. A common observational method it time-sampling, in which the total observation time is subdivided into equal time units. Observers usually record whether or not the target behavior occurred during each time unit. The level of inter-observer agreement is then assessed by dividing the number of intervals in which both observers recorded that the target behavior occurred by that same number plus the number of intervals in which only one of the observers recorded the occurrence of the behavior. The formula derived from this procedure is: Inter-observer agreement = Number of agreement intervals/number of agreement intervals plus number of disagreement intervals. While

> . . . samples of observer agreement . . . provide the basis for assertions that recordings are the reliable product of well specified recording procedures, rather than the idiosynchratic judgments of several observers. . .(Kent & Foster, 1977, p. 296)

available evidence indicates that under certain circumstances inter-observer agreemnnt may be spuriously inflated and not representative of agreement during the entire observation session.

Reid (1970) was the first researcher to demonstrate that inter-observer agreement may be considerably higher when the observers know that inter-observer agreement checks are taking place than when they are not aware that their agreement is being assessed. In a more detailed study, Taplin and Reid (1973) compared the effects of three instructional sets on inter-observer agreement. One group of observers was told their observations would not be subject to assessment of inter-observer agreement; a second group was told to expect agreement to be assessed randomly at any point during the study; a third group was told their observations would be subject to spot-checking of agreement, and that they would be aware when a spot-check was occurring. In fact, agreement levels were continually monitored for all groups. The results indicated that the "random-check" group maintained the highest overall agreement of the three groups. Moreover, the "spot-check" group evidenced the lowest agreement on all days except on the days when a spot-check occurred. The "no-check" group's agreement levels were consistently lower than those of the "random-check" group, and, except for days on which a spot-check occurred, were consistently higher than those of the "spot-check" group.

Inter-observer agreement may also be spuriously high when observers know with whom they will be comparing their data during assessment of inter-observer agreement. Romanczyk, Kent, Diament, and O'Leary (1973) purposely trained two

assessors to interpret four of the nine disruptive behaviors on the O'Leary Code in a consistent but different way, so that one assessor would consistently obtain higher frequency counts for two behaviors. The authors examined agreement between each of five trained observers and each of the two assessors under two experimental conditions, overt and covert assessment. During overt assessment, each observer was informed that agreement with one of the assessors would be checked at the end of the session. During covert assessment, observers were not informed that agreement would be assessed. Actually, throughout both conditions both assessors recorded data with the observer whose agreement was being assessed.

Both assessors had calculated inter-observer agreement with each observer an average of four times during training prior to the experimental phase of the study. Each observer, then, had four opportunities to note how each assessor interpreted the behavioral code.

A comparison of each observer's data with each assessor across all experimental sessions, however, indicates that for the four modified behavior codes, observers matched the records of the known assessor (.77 agreement) to a greater extent than they did the unknown assessor (.53 agreement). The authors concluded, then, that when observers know the identity of the person with whom their level of agreement is assessed, they can often modify their recordings to match that of the assessor.

Observer Drift

The studies cited above describe some of the conditions under which levels of inter-observer agreement may be spuriously inflated, and thus unrepresentative of the level of agreement reported when assessment of agreement is performed continuously. The Romanczyk et al. (1973) study typifies a phenomenon mentioned earlier, namely, observer drift. Johnson and Bolstad (1973) suggest that this phenomenon occurs when observer pairs or groups begin to differ in their interpretation of a behavioral code, and modify their recordings accordingly so as to maintain high levels of inter-observer agreement. When high inter-observer agreement is achieved but individual observer accuracy is sacrificed, the result may be a decline in the overall accuracy of the data collected by the observer pair. The Kass and O'Leary (Note 1) study demonstrated one way that observer drift may affect the internal validity of a study. Observers were trained in separate groups and then were assigned by groups to different levels of the independent variable (different expectancies). The data produced could not be attributed solely to expectancies because different expectancies were confounded with different groups of observers. Other studies have also examined observer drift.

Kent, O'Leary, Diament, and Dietz (1974) trained twenty observers as a group to use the O'Leary Code to record children's behavior from videotapes. Then they paired observers

and had each pair practice recording together for five hours before assigning pairs to different expectancy conditions. The authors included five factors in the design: expectancy, treatment condition, observation session, target child, and observer pair. An analysis of variance revealed that expectancy did not affect the observers' behavioral recordings. However, observer pairs within expectancy conditions did differ significantly on three of the nine behavioral categories (Playing, Orienting, and Noise). In addition, the effects of the different recording of observer pairs interacted with several other factors in the design. For the category Playing, for example, all interactions involving observer pairs were significant. For the composite Total Disruptive Behavior score, sources involving observer pair accounted for 17% of the variance. Five percent of this was due to observer pair only. The authors conclude from their data that "it seems unwise, therefore, to confound individual observers or groups of observers with experimental conditions in studies employing behavioral recording" (p. 779).

Studies of observer training variables have offered a chance to examine the effects of observer drift. DeMaster, Reid, and Twentyman (1977) trained 28 observers as a group to record behavior from videotapes, then paired observers and assigned each pair to one of three feedback conditions. Observers in Group One were given feedback regarding the accuracy of their recordings. Observer accuracy refers to a

comparison between an observer's record and a standard, or record which is assumed to represent the actual occurrence of the behavior being observed. This is quite different from inter-observer agreement, where an observer's record is compared to that of another observer, whose record is not necessarily more representative of the actual occurrence of the behavior than that of the other observer. Observers in Group Two received feedback regarding inter-observer agreement. Observers in Group Three received no feedback regarding their observational recordings. The authors found that agreement within observer pairs was consistently higher than between members of different pairs. Inter-observer agreement within observer pairs was also higher than observer accuracy for all three groups. However, observers in Group One, who had been given feedback regarding their accuracy (by comparing their records to a previously established standard), were more accurate than Group Two observers (who only discussed inter-observer agreement), who were in turn more accurate than observers in Group Three. Johnson and Bolstad use these findings to suggest the use of videotaped material and standards of comparison as a means of enhancing observers' accuracy.

Wildman, Erickson, and Kent (1975) also reported effects of different types of observer training on inter-observer agreement. The experimenters trained 16 undergraduate students to record the behavior of nursery school children.

The observers worked in pairs; one-half of the observer pairs were trained by a single graduate student trainer; the other half trained by themselves. All observer pairs then coded behavior from four 10-minute videotapes.

The authors reported several interesting findings: first, for all observer pairs, overtly-assessed agreement levels were higher than covertly-assessed levels (corroborating Reid's 1970 findings); second, although group differences in agreement were not found, the authors did find group differences in the behavioral observations (the group trained by the graduate student recorded a higher frequency of behaviors and showed less variability about the mean frequency than did the group which trained itself); finally, within-pair agreement levels were higher than agreement levels calculated between different pairs of observers.

Observer "Cheating"

The research discussed thus far has described some of the variables affecting the representativeness of: behavioral recordings versus qualitative ratings; and the levels of inter-observer agreement obtained and reported by pairs of observers. It is unclear to what extent the aforementioned studies involved intentional altering of data or carelessness in recording behavior or in computing agreement.

Rosenthal (1966) states, however, that intentional data fabrication is present in psychological research, especially to the extent that experimenters consider it more important

to produce "desirable" data than to adhere to strict scientific procedures. This attitude may result from demands, implicit or explicit, from investigators in the study, as well as from the experimenter's perception of the consequences of adopting these alternative strategies. In discussing "Investigator Fudging," Barber (1976) suggests that competition for prestige among scientists may provide the motivation for fudging or biasing their data. This pressure may be transmitted to the experimenter, albeit unintentionally, with dramatic results. A study by Rosenthal and Lawson (1964) revealed that students clearly fabricated data in the context of an animal learning experiment. Azrin, Holz, Ulrich, and Goldiamond (1961) found similar instances of intentional data fabrication in the context of replicating a verbal conditioning experiment.

Weber and Cook (1962) categorized subjects in psychological experiments into four major categories: "good" subjects are those who are aware of the experimental hypothesis and attempt to produce data to confirm it; "negativistic" subjects are those who are also aware of the experimental hypothesis, but who attempt to produce data to disaffirm it; "apprehensive" subjects, those who make the responses they perceive will make them "look good" to the experimenter; and "faithful" subjects, who attempt to remain as objective as possible and adhere to scientific procedure.

This schema may also apply to experimental assistants. "Good" experimental assistants may be those who know the

experimental hypothesis and who without awareness attempt to produce data to confirm those hypotheses; that is, they unwittingly produce biased data. "Apprehensive" experimental assistants may attempt to produce data that make them appear to be conscientiously following experimental procedures; that is, they may cheat to produce data that seem to have been carefully collected. According to Kent and Foster (1977), some observers may actually alter their observational records while computing inter-observer agreement to increase the level of agreement reported; or, they may make computational errors when they calculate agreement, also to increase agreement levels they report. Two variables which may enhance such cheating are absence of supervision by the experimenter and permission for observers to communicate during times when observers calculate inter-observer agreement (Kent & Foster, 1977).

O'Leary and Kent (1973) present evidence that observer pairs produce higher levels of agreement when calculating agreement in the experimenter's absence than in his presence. Twelve observers were trained to record behavior from videotapes, and to calculate agreement in the experimenter's presence. During the last six training days, however, the experimenter was called from the room following one of the two observation sessions. A comparison of the observers' records revealed that observers reached an overall level of .66 agreement in the experimenter's absence, and only .55 in his presence.

Observers may also make errors in computing agreement which result in spuriously high agreement levels being reported. These errors may include incorrect addition or division, or transferring incorrect numbers from one area of the data summary sheet to another. Regardless of the type of error, the result is often the reporting of an incorrect level of inter-observer agreement.

O'Leary and Kent (1973), for example, supervised the behavioral recordings of ten observers without closely monitoring their calculations of inter-observer agreement. When experimenters re-examined the observers' calculations, they discovered errors which had inflated the level of agreement by eight points. Observers had reported average agreement levels of .66; the accurate level of inter-observer agreement re-calculated by the experimenters was .58.

Kent et al. (1974) also discovered calculation errors in calculations of inter-observer agreement of two groups of observers. The observers' calculations produced agreement levels of .76 and .73; the experimenters used the same data to produce agreement levels of .67 and .70.

The O'Leary and Kent (1973) and Kent et al. (1974) studies investigated errors that observers make when they calculate their own levels of inter-observer agreement. Rusch, Walker, and Greenwood (1975) examined errors made by experimenters when they summarize data collected by observers. Two research assistants (experimenters) were asked to summarize

the observational data collected during an experimental classroom interventinn program. The data summaries included the calculation of response rates and durations for each of six children, and mean rates and durations for all children for each observation session. Two other research assistants and a computer staff also checked the data and provided summary standards against which the experimenters' summaries could be compared. The authors noted some discrepancies between the experimenters' summaries and the standards; however, in most cases the discrepancies were slight, and showed no consistent tendency to be in the direction of the experimental hypothesis. The authors concluded that the experimenters' summaries would have yielded the same conclusions regarding the effectiveness of the classroom intervention program as the standard summaries would have. It would have been interesting to also have asked observers to summarize their own data, to examine the possible effects on calculation errors.

Statement of Purpose

Rosenthal (1963, 1966) and other social psychologists provided the early impetus for research into the effects that experimenters may have on the outcome of their experiments. Despite the failure of researchers to replicate many of Rosenthal's findings, and despite the compelling criticisms of much of his research (Barber & Silver, 1968), a great

deal of concern has been generated regarding the potential for experimenter bias effects.

Drawing an analogy between experimenters and behavioral observers, behavioral researchers have investigated the conditions under which observer biases may occur. The best evidence to date indicates that the effects of antecedents (expectancies and demand characteristics) are more pronounced on observers' global ratings of the behavior after they observe it than on their behavioral recordings of the behavior as they observe it. Consequent events (for example, providing evaluative feedback to observers about their recordings), though, do appear to bias behavioral recordings, making the behavioral record unrepresentative of the actual occurrence of the behavior.

Research has also provided some preliminary findings regarding conditions under which the levels of inter-observer agreement by observers during specific portions of data collection may not be representative of the level of agreement for the entire observational period. For example, when observers know when inter-observer agreement is being assessed, or when they know the identity of the agreement assessor, they may produce spuriously high levels of agreement. Observer pairs may also "drift" in their application of the behavioral code, resulting in high inter-observer agreement but low accuracy. Finally, observers may intentionally fabricate data, or make mistakes in calculating agreement, resulting in the

reporting of data that appear to be consistent with experimental rules or procedures, but are not representative of the actual occurrence of the behavior or the actual level of agreement.

Because behavioral research and therapy rely heavily on data collected in natural settings by human observers, the importance of studying variables which may affect the quality of these data cannot be understated. If the data are to be clearly interpretable and have implications beyond the specific research or therapy setting, researchers must be able to take advantage of procedures which will most likely assure that the data collected represent the actual occurrence of the behavior observed.

The present study examined the effects of two variables on the quality of observational data and on computations of inter-observer agreement. The first variable was a particular instructional set. Observers in field experimental settings are generally given a number of instructions to follow. Usually they are told to act as a mechanical recording device, and not to look at the other observer to see if they should record a particular behavior. In short, there is an implicit demand for the observer to record behavior as objectively and carefully as possible, independent of the behavior of the other observers. In addition to the implicit demand for objectivity in observation, there is also an implicit demand to summarize the data carefully.

On the other hand, observers are told that high levels of inter-observer agreement are essential in order to separate treatment effects from random error in the data. Because most experiments do not permit the actual occurrence of the behavior to be assessed, inter-observer agreement is accepted as the basis on which the most nearly accurate assessment of the occurrence of the behavior is made.

This study manipulated the instructional set with regard to the demands made on the subject. Half the subjects were instructed to try to reach a certain level of inter-observer agreement (.85), while the other half were instructed to make their own observational recordings and agreement calculations as objectively and carefully as possible.

The second variable related to the procedures for calculating inter-observer agreement. O'Leary and Kent (1973) and Kent et al. (1974) demonstrated that when observers are not supervised closely, they may make errors in calculating agreement between their own and their partner's records. Rusch et al. (1975) found insignificant calculation errors made by experimenters summarizing the data collected by others; however, the experimenters in this study did not collect and summarize their own data.

Subjects in this study calculated agreement levels on two sets of observational records following each of eight observation sessions: one set were the actual recordings of the observer pair; the other set were constructed by the

experimenter, although subjects were told the data were collected by other observers. The order of calculation of agreement levels for the two sets of data were counterbalanced to control for sequence effects.

Observers were trained as a single group before being assigned a partner. Observer pairs were randomly assigned to one of the instructions groups. The pair recorded from videotapes the occurrence of three behavioral responses of one of two second-grade classroom children. Each child was observed half of the time, and the order in which the children are observed was counterbalanced to control for sequence effects. However, the target child observed was not a factor in the design, since both children were males, and the behavioral definitions were the same for both of them. Thus the design was a 2 (instructions) x 2 (data sets) x 3 (target behaviors) x 8 (experimental sessions) Mixed design. Instructions was the only between-subjects factor.

Four dependent measures were studied: the accuracy of each observer's record (obtained by comparing each observer's record with an already established standard); the actual agreement levels obtained by observer pairs (obtained by having the experimenter independently calculate agreement for the records of the observer pair); the difference between observer accuracy and inter-observer agreement for each subject; and the difference between subject-calculated and experimenter-calculated agreement figures.

Experimental Hypotheses

1. Regarding observer accuracy: the group of observers instructed to observe the tapes and calculate inter-observer agreement as carefully as possible was expected to have obtained accuracy levels higher than the group instructed to try to achieve a high level of inter-observer agreement when accuracy is summed over all sessions and target behaviors. This would demonstrate that the latter group had drifted in their application of the behavioral code in order to maintain high agreement levels.

2. Regarding inter-observer agreement: actual levels of inter-observer agreement obtained by observer pairs instructed to try to reach a specific high level of agreement were expected to have been significantly higher than the actual levels of inter-observer agreement obtained by pairs of observers instructed to observe and record behavior from the videotapes and perform calculations carefully. Since high levels of accuracy obtained by a pair of observers necessarily means that their agreement level is also high, the degree to which the groups differ should have depended in part on the level of accuracy obtained by members of the same pair in each group.

3. Regarding the difference between observer accuracy and inter-observer agreement: whether or not group differences on this variable were significant depended on the magnitude and direction of group differences on the first two

dependent measures. For the group instructed to try to reach high agreement scores, agreement was expected to have exceeded accuracy, yielding a positive difference score. For the group instructed to observe carefully, the difference score was expected to have been either a smaller positive number or, more likely, a negative number.

4. Regarding the difference between subject-calculated and experimenter-calculated agreement scores, a Groups x Data sets interaction was predicted. The high inter-observer agreement group was expected to spuriously inflate their own agreement figures (yielding a positive difference score) and spuriously deflate agreement figures of "other" observers (yielding a negative difference score). The difference scores for the observers instructed to observe carefully were expected to be smaller; also, directional errors similar to those predicted for the other group were not expected.

CHAPTER II

METHOD

Subject Selection

Nineteen undergraduates from an introductory psychology course received course credits for participating in a study of children's classroom behavior. The experimenter asked the students to complete ten selected arithmetic problems from the Wide Range Achievement Test (Jastak, Bijou, & Jastak, 1961) (see Appendix A) prior to the training phase of the experiment. This was done to assure that arithmetic skill was roughly equivalent in the experimental groups.

The students were then trained to code children's classroom behavior from videotapes. Training continued until 16 students reached a pre-established criterion. These 16 students (14 females, 2 males) became the experimental subjects. Subjects were first assigned to a partner according to the following factors: WRAT scores (four subjects who scored below 80% were assigned partners who scored 80% or above); number of hours of training (subjects who were trained only the minimum of one hour were assigned partners who were trained the maximum of three hours); training session attendance (an attempt was made to assign subjects partners with whom they had little or no contact during training); and subjects' schedules. Pairs of subjects were then randomly

assigned to one of the two experimental groups (except that subject pairs which included a subject who had scored below 80% on the WRAT were divided equally between the two groups). Table 1 presents a description of the experimental groups with respect to the above factors. There were eight subjects (four observer pairs) in each group.

Design

The experimental design consisted of a 2 (instructions) x 2 (data sets on which subjects performed arithmetic calculations) x 3 (target behaviors) x 8 (sessions or trials) Mixed design. Instructions, the between-subjects factor, were either to perform the observational recording and calculation tasks as carefully as possible, or try to reach a high level of inter-observer agreement. (The latter set of instructions indicated that .85 was a minimally acceptable level of agreement to reach.) The data sets were either the observer pair's own data for a session or the data they were told were collected by two other observers, but which were actually contrived by the experimenter. (This factor was counterbalanced to control for sequence effects; see Appendix B). The three target behaviors were Playing, Vocalizing, and Orienting. A trial was completed when an observer pair had recorded the occurrence of each of the three target behaviors from a 10-minute videotape and had calculated inter-observer agreement for each behavior on each of the two data sets.

Apparatus

A training manual (see Appendix C) was developed and given to each subject prior to training. It included a brief rationale for the study; the behavioral code; training and experimental phase procedures (including an explanation of the time-sampling observational method, coding procedures, and the procedure to follow when calculating inter-observer agreement); and a hypothetical set of data showing one way to correctly calculate agreement figures.

Videotapes were selected from a collection of tapes developed by researchers at the State University of New York at Stony Brook. The Psychology Department of the University of North Carolina at Greensboro has copies of these tapes. Each one consists of two 12½ minute samples of the classroom behavior of two, 8-year-old second grade males. An audio signal indicates when each sample begins and ends.

Several videotapes were shown during the training phase of the experiment. A five-minute portion of one tape was presented at the first training session. The first minute of the sample was used merely to expose observers to the format of the tapes and to identify the children on the tapes. During the next two minutes the experimenter identified examples of the target behaviors. During the final two minutes, observers were encouraged to comment verbally when they detected occurrences of any of the target behaviors by a specific child.

Four other 5-minute videotape segments were used during the training period as "probe tapes"; that is, observers actually recorded their own observations of the behavior of a particular child. Their data were later scored for accuracy; if their accuracy scores met criteria and if they had demonstrated the ability to correctly calculate one set of three inter-observer agreement scores, they were considered trained and ready for the experimental phase of the study.

Eight other 10-minute samples were selected for the experimental phase of the study.

The researchers at the State University of New York at Stony Brook also developed protocols for the videotapes. A number of well-trained observers, using the Classroom Behavior Code to record simultaneously the occurrence of nine target behaviors, repeatedly viewed the tapes until a consensus of agreement regarding the occurrence of the target behaviors was reached. These researchers have assumed that the result of many observers, each highly skilled at applying the behavioral code, reviewing the tapes many times, is a valid method for assessing the actual occurrence of behavior on the tapes. Thus, the protocols were used as the standard for assessing each subject's observational accuracy.

The experimenter developed data sheets equipped with carbon so subjects could duplicate their data for each trial. This allowed each member of an observer pair to calculate agreement simultaneously with his or her partner. The experimenter also developed eight contrived data sets for which

each subject calculated agreement levels during each trial. One data sheet represented the protocol for the videotape for that trial; the other data sheet was constructed so that correct agreement levels for the data set would vary around .85, the agreement level which half the subjects were instructed to try to reach.

A questionnaire (see Appendix D) was developed and filled out by each subject at the end of the last trial. The questionnaire was designed to determine if the subject could state the instructions he or she was given prior to odd-numbered trials.

A Panasonic videotape recorder (model NV3020) and 19-inch diagonal screen monitor (model AN69V) were used to present the tapes. An audio signal heard on a Sharp cassette tape recorder cued subjects to correct observation and recording intervals.

Subjects were trained at the McNutt Media Center of the University of North Carolina at Greensboro. The experimental phase of the study was conducted in Room 301, Nursing Building, also on the University of North Carolina at Greensboro campus.

Coding Procedures and Calculation of Inter-Observer Agreement

Subjects coded the occurrence of three target behaviors by the children: Playing (using hands to manipulate own or other's property, in a manner incompatible with learning);

Vocalizing (emitting from the mouth any "non-permitted audible response); and Orienting (turning the head more than 90° from the point of reference while seated). The complete behavioral code is found in Appendix C. The experimenter used two criteria in his selection of target behaviors from the nine included by the Classroom Behavior Code. First, high-frequency (based on the protocols) behaviors were sought, since it was assumed that high frequency might increase the possibility subjects would make calculation errors (since they would be counting and dividing larger numbers). Second, behaviors which were the easiest to discriminate were sought, in order to facilitate observer training. Previously reported levels of agreement for the behaviors contained in the code were used to make this latter decision (Kent, O'Leary, Colleti, & Drabman, Note 6).

The same time-sampling procedure used to develop the protocols was used in this study. An audio signal cued subjects to observe the child for 20 seconds. At the end of 20 seconds, another cue signaled them to record their observations during the next ten seconds. This procedure yielded, for a 10-minute tape, 20 intervals of 30 seconds each.

After they made observational recordings from a 10-minute videotape, subjects calculated inter-observer agreement for each of two data sets: the observer pair's own data and the contrived data set for that session. Thus, each subject reported six agreement figures (three target behaviors x two data sets) for each session.

Subjects used the Exact Agreements formula (Repp, Dietz, Boles, Dietz, & Repp, 1976) to calculate inter-observer agreement. The formula is: Agreement Intervals/Agreement Intervals + Disagreement Intervals. An agreement occurs in our interval when both observers record that the behavior occurred. A disagreement occurs when only one of the two observers record that the behavior occurred. Intervals when neither observer records that the behavior occurred are omitted from the calculations.

Training Procedure

Students who participated in the study received a training manual to read prior to the first meeting. At this meeting, the experimenter re-read the rationale for the study. After signing consent forms, the students completed the WRAT arithmetic items. The experimenter then briefly reviewed the rest of the training manual.

The students then saw the first 5-minute training tape. The first minute of this tape was used to expose the students to the format of the tapes. During the next two minutes the experimenter pointed out examples of the target behaviors emitted by one of the children. During the last two minutes the tape was stopped after each interval, and the students discussed the behavior they had seen. They were told that the experimenter's observations for the 5-minute tape segment were recorded on one of the data sheets in their manual. These data were actually drawn from the protocol.

The experimenter then referred the students to the second completed data sheet in the manual; this sheet demonstrated one way to correctly calculate inter-observer agreement. Students were encouraged, however, to develop their own system, making sure to enter their final figures in the appropriate boxes.

After answering procedural questions, the experimenter handed out blank data sheets. Each student collected data from another 5-minute sample of tape. The experimenter answered any questions regarding the tape content, then asked the students to exchange a copy of their data with the person next to them. Each student then practiced calculating agreement. The experimenter walked around the room and answered any questions related to this procedure. The first training session lasted one hour.

The experimenter calculated accuracy levels for each subject's data from the first "probe" tape by comparing the subject's data sheet to the protocol for that segment of tape. The formula used was the same as that used for calculating inter-observer agreement (Agreements/Agreements + Disagreements). Subjects whose accuracy levels for two of the three target behaviors was .80 or higher and whose inter-observer agreement calculations were done correctly were considered trained and ready for the experimental phase of the study.

Subsequent training sessions were held until 16 subjects reached training criteria. At these training sessions the

experimenter re-read the behavioral coding system and answered questions about it. Subjects then collected data from other 5-minute taped segments. Since two target subjects appeared on all tapes, each of the four probe tapes could be used twice during training. (See Table 1 for a description of the average amount of training subjects in each group received.) The content of each tape was discussed at its conclusion. The experimenter used the tape protocols to answer questions about the occurrence of specifying behaviors during certain intervals. Also, subjects continued to practice calculating agreement on the data collected by other observers as well as on their own data. Sixteen subjects reached training criterion by the end of the third 1-hour training session.

Experimental Procedures

On each day the observer pair entered the laboratory in Room 301, Nursing Building (on the UNC-G campus) and were seated about four feet apart, and about seven feet from the monitor. On a blackboard behind the monitor were a brief version of the behavioral code and the formula for calculating agreement. After answering preliminary questions, the experimenter read a set of instructions to the subjects (see Appendix E), depending on the group to which they were assigned. The observer pair then collected data (on a child specified by the experimenter) from a ten-minute segment

of videotape. The experimenter remained in the laboratory during this time and collected the observers' data sheets at the end of 10 minutes.

After collecting their data sheets, the experimenter gave each subject a set of data and asked the subject to calculate agreement figures for the three behaviors. After completing this task, the subject was given a second data set on which to perform the same task. The data sets were either the observer pair's own data or were contrived by the experimenter to look like data collected by other observers. The order of presentation of the two data sets was counterbalanced to control for sequence effects.

Subjects used the same formula as that used during training for calculating agreement (Agreements/Agreements + Disagreements). Each subject produced six final agreement figures (two data sets times three target behaviors).

Completion of agreement calculations on the second data set marked the end of a trial. Each observer pair completed four trials on each of the two days. Instructions were read to subjects at the beginning of odd-numbered trials.

Dependent Measures

Four dependent measures were examined. The first three discussed relate to the observational task the subjects performed. The fourth relates to the calculation tasks they performed after each observational session. First, inter-observer agreement (the degree to which two observers making

simultaneous observation agree that a particular behavior occurred) was assessed for each observer pair. The experimenter calculated agreement figures for each set of data collected by each observer pair during the study.

Second, each subject's observational accuracy was assessed. The experimenter compared the subject's observational data for each trial to the protocol of the videotape for that trial. Since the protocol is assumed to best represent the behavior actually occurring on the tape, the precision of each subject's observations may be examined.

Third, the difference between each subject's agreement with his or her partner (inter-observer agreement) and his or her agreement with the protocol (accuracy or precision) was assessed. The experimenter subtracted each accuracy value, obtained above, from the corresponding agreement figure. This measure attempted to show the relationship between the first two dependent measures.

Fourth, the difference between subject-calculated inter-observer agreement and the correct figures was assessed. The experimenter assessed this difference for both sets of data on which subjects calculated agreement. For the subject's own data, the experimenter subtracted the figure previously calculated by him (dependent measure number one) from the final figure reported by the subject on the data sheet. For the contrived data the correct agreement figures were pre-established (see Appendix F); each was subtracted from the

final figure on the data sheet on which the subject performed the calculations. This is a measure of the magnitude and direction of errors made in calculating and reporting inter-observer agreement.

In addition, the experimenter performed a quality control check of his own work. He re-checked 20% of his own work; this meant checking the accuracy of 346 numbers of the 1,728 total calculations he performed. No errors in this sample of the experimenter's own work were found after the second check.

CHAPTER III

RESULTS

Data Preparations

At the end of the study each subject had produced the following data: eight observational records (from four observational sessions per day over two days) of the occurrence of each of three behaviors from eight 10-minute segments of videotape (each divided into 20 intervals); and a total of 48 agreement figures (three behaviors times two data sets times eight trials), half of which are from the subject's own data, and half from contrived data. Each subject also completed a questionnaire.

The experimenter had prepared the following data: eight observational records chosen from the protocols of the experimental tapes; and the correct agreement levels for the eight sets of contrived data (24 correct figures, based on eight data sets times three target behaviors).

To compile the data the experimenter first used the protocols from the experimental tapes to calculate each subject's accuracy levels for each trial. The standard formula (Agreements/Agreements + Disagreements) was used in the calculations to produce a grand total of 384 accuracy figures (three behaviors times eight trials times 16 subjects), each varying from zero (when no agreements between two data sheets are found) to one (when no disagreements are found).

The experimenter next re-calculated agreement levels for each observer pair's own data for each session (each subject had already performed this task during each session), to assess the correct agreement levels for each observer pair. This procedure yielded 192 agreement figures (three behaviors times eight trials times eight observer pairs), each again varying from zero to one.

Next the experimenter subtracted each subject's accuracy figure for each behavior over each trial from the agreement figure for the same behavior and trial (each agreement score was used twice, since agreement scores were the same for both members of an observer pair). This procedure produced 384 difference scores, each varying from one (when agreement equalled one and accuracy equalled zero) to minus one (when agreement equalled zero and accuracy equalled one).

The last step in compiling the data was to determine the difference between subject-calculated inter-observer agreement and experimenter-calculated ("correct") agreements for both data sets. For the contrived data, correct agreement figures were pre-established. Each figure was subtracted from the figure reported by each subject for that behavior for that trial. This yielded 384 difference scores, each varying from .30 to -.94 (correct agreement figures varied from .94 to .70; subject-calculated figures varied from one to zero; thus, 1-.70=.30 and 0-.94=-.94).

For the subject's own data the procedure was similar. The experimenter had already calculated each subject's correct

agreement figures. Each correct figure was subtracted from the figure reported by the subject for that behavior and trial. The result was 384 difference scores, each varying from one to minus one (since the range of subject-calculated and correct agreement scores varied from zero to one).

Questionnaire Data

Each subject completed a questionnaire at the end of the final trial of the experiment (see Appendix D). The primary purpose for this was to ascertain whether or not subjects could recall the content of the instructions given to them prior to each odd-numbered trial.

An analysis of the questionnaire data revealed an interesting finding. Table 1 presents the results of the analysis of the questionnaire data. All the subjects instructed to make their observational recordings and calculations as carefully as possible correctly recalled these instructions. Three of eight (37%) subjects instructed to try to reach a level of .85 agreement with their partner correctly recalled these instructions; however, five of eight (63%) recalled that they were instructed to try to reach .85 agreement levels with their partner and to make their recordings and calculations as carefully as possible.

One explanation for these findings lies in the questionnaire itself. The question asked of subjects was a multiple-choice item which allowed subjects to circle the response

Table 1

Group Descriptions

Descriptor	High Agreement Group	Careful Group
Sex	7 female; 1 male	7 female; 1 male
Average age	24 (range: 20-39)	30 (range: 21-51)
Average WRAT Score	86% (range: 70-100)	80% (range: 50-90)
Average hours Training per Subject	2.13 (range: 1-3)	2.38 (range: 1-3)
Questionnaire Data: Percent correctly recalling instructions	63% (5 of 8)	100% (8 of 8)

which indicated that they had received both instructions to try to reach .85 agreement levels with their partner and to make their recordings and calculations as carefully as possible. One might speculate that the results might have been different if: subjects had been required to briefly write the main points of the instructions, without prompts; or, if subjects had been forced to choose only one set of instructions.

The other explanation suggests that those subjects who responded that they had received instructions to record carefully as well as reach .85 agreement with their partner operated under both instructional sets. This means that perhaps the demand for high agreement was tempered somewhat for some subjects. This might account for the finding that instructions did not significantly affect observer accuracy or inter-observer agreement separately.

Overview of Statistical Analyses

An analysis of variance was performed on each of the four dependent measures. For accuracy (N=16), agreement (N=8), and agreement minus accuracy (N=16), the analysis consisted of a 2 (instructions) times 3 (behaviors) times 8 (trials) ANOVA. For subject-calculated agreement minus experimenter-calculated agreement (N=16), the analysis consisted of a 2 (instructions) times 2 (data sets) times 3 (behaviors) ANOVA. Scheffe' post-hoc comparisons among means were calculated for significant main effects for behaviors and trials.

Calculation Errors

Table 2 gives the results of the analysis of variance for calculation errors (subject-calculated minus "correct" or experimenter-calculated agreement figures). The main effect of different data sets was significant, F (1, 14) = 9.64, $p < .01$. Observers, as a single group, spuriously inflated their own agreement scores by .04, and spuriously deflated the agreement scores from the contrived data by .07. No other signifcant effects were found.

It was predicted that a significant Groups x Data sets interaction would be found. Instructions were expected to affect the high agreement demand group, resulting in their spuriously inflating their own agreement levels, and spuriously deflating agreement levels from the contrived data. Agreement levels for both sets of data calculated by observers in the careful group were expected to be near zero and nondirectional. However, the pattern of results for the two groups was similar enough to contribute to the significant main effect of data sets.

Difference Scores (Agreement minus Accuracy)

Table 3 presents the results of the analysis of variance of the difference scores resulting from subtracting observer accuracy levels from agreement levels. As predicted, the main effect of groups is significant, F (1, 14) = 5.25, $p < .05$. For the high agremment demand group, overall agreement levels exceeded accuracy levels by five points. For the careful

Table 2

Summary of Analysis of Variance of Observer Calculation Errors (Subject-calculated Agreement--Experimenter-calculated Agreement)

Source	SS	df	MS	F
Group (G)	.15	1	.15	2.224
Data set (D)	.299	1	.299	9.637*
Behavior (B)	.053	2	.026	2.178
Error S(G)	.947	14	.068	
G x D	.001	1	.001	.019
G x B	.006	2	.003	.23
D x B	.046	2	.023	.584
Error SD(G)	.435	14	.031	
Error SB(G)	.342	28	.012	
G x D x B	.07	2	.035	.9
Error SDB(G)	1.096	28	.039	

* $p < .01$

Table 3

Summary of Analysis of Variance of Difference Score
(Inter-observer Agreement--Observer Accuracy)

Source	SS	df	MS	F
Group (G)	.541	1	.541	5.249*
Behavior (B)	.409	2	.204	.841
Trial (T)	.641	7	.092	2.747**
Error S(G)	1.444	14	.103	
G x B	.035	2	.018	.728
G x T	.246	7	.035	1.053
B x T	1.074	14	.077	2.411***
Error SB(G)	.681	28	.024	
Error ST(G)	3.268	98	.033	
G x B x T	.784	14	.056	1.761*
Error SBT(G)	6.238	196	.032	

* $p < .05$

** $p < .025$

*** $p < .01$

group, the relationship was reversed: accuracy was two points higher than agreement.

The main effect of trials was also significant, F (7, 98) = 2.75, $p < .025$. Mean difference scores (agreement minus accuracy) for Trials 1 through 8 were .01, -.05, .09, .002, -.007, .04, -.04, and -.01, respectively. It can be seen that agreement levels exceeded accuracy levels by the greatest amount on Trial 3 (.09). Accuracy levels exceeded agreement levels by the greatest amount on Trial 2 (-.05).

The Behaviors x Trials interaction was significant, F (14, 196) = 2.41, $p < .01$. The greatest difference in the agreement/accuracy relationship was found between Playing and Vocalizing on Trial 1. For Playing, accuracy was nine points higher than agreement, while for Vocalizing, agreement was ten points higher than accuracy.

The Groups x Behaviors x Trials interaction was also significant, F (14, 196) = 1.76, $p < .05$.

Observer Accuracy

Table 4 presents the results of the analysis of variance of observer accuracy levels. The main effect of ehaviors is significant, F (2, 28) = 17.52, $p < .005$. Observer accuracy for Playing (.52) was ten points lower than for Vocalizing (.62). This may have been due to observers' difficulty in distinguishing between a child's manipulating an object in a manner "incompatible with learning" and his manipulating an object as a part of his assignment. Observer accuracy for Orienting was .60.

Table 4

Summary of Analysis of Variance of Observer Accuracy

Source	SS	df	MS	F
Group (G)	.258	1	.258	3.256
Behavior (B)	.813	2	.406	17.522*
Trial (T)	2.89	7	.413	18.53*
Error S(G)	1.11	14	.079	
G x B	.017	2	.008	.364
G x T	.291	7	.042	1.864
B x T	5.241	14	.374	17.626*
Error SB(G)	.649	28	.023	
Error ST(G)	2.183	98	.022	
G x B x T	.413	14	.295	1.39
Error SBT(G)	4.163	196	.021	

* $p < .005$

The main effect of trials was also significant, F (7, 98) = 18.53, $p < .005$. Mean accuracy levels for Trials 1 through 8 were .65, .55, .50, .71, .53, .50, .50, and .71, respectively. A Scheffe post-hoc comparison among means reveals that observer accuracy for Trials 4 and 8 was significantly higher than for Trials 3, 6, and 7, $p < .05$. These results will be discussed more fully in the next chapter.

The analysis revealed a significant Behaviors x Trials interaction, F (14, 196) = 17.63, $p < .005$. The greatest difference in observer accuracy occurred in Trial 7, where observer accuracy for Playing was only .17, while accuracy for Vocalizing was .77, a difference of 60 points.

It was predicted that observer group differences would be found. The group instructed to observe carefully and independently was expected to reach an overall higher level of accuracy than the group instructed to try to reach high agreement with their partners. This would reflect a greater amount of drift (or unique application of the code) by the latter group.

The results of the ANOVA indicate that the difference in overall accuracy between the two groups approached, but was not, significant, F (1, 14) = 3.26, $p < .10$. The careful group achieved an overall accuracy level of .61; the high agreement demand group, .56.

Inter-observer Agreement

Table 5 presents the analysis of variance of inter-observer agreement levels. The pattern and interpretation of results is similar to the results for accuracy levels.

The main effect of behaviors was significant, F (2, 12) = 13.11, $p < .005$. The average agreement level for Playing (.53) was 11 points lower than for Vocalizing (.64). The agreement level for Orienting was .62.

The main effect of trials was significant, F (7, 42) = 4.54, $p < .005$. The average agreement levels for Trials 1 through 8 were .66, .50, .64, .71, .52, .54, respectively. As can be seen, the largest difference in agreement levels was between Trial 2 (.50) and Trial 4 (.71), a difference of 21 points.

The Behaviors x Trials interaction was also significant, F (14, 84) = 7.04, $p < .005$. The greatest difference in agreement levels occurred on Trial 7, where agreement for Playing was only .22, while agreement for Vocalizing was .83, a difference of 61 points.

It was predicted that observer "drift" might be reflected by significantly higher agreement levels among observer pairs instructed to try to obtain higher agreement levels than among those in the group instructed to observe carefully and independently. Contrary to prediction, the main effect for groups did not approach significance. The high agreement demand group achieved an overall level of .62

Table 5

Summary of Analysis of Variance of Inter-observer Agreement

Source	SS	df	MS	F
Group (G)	.047	1	.047	.243
Behavior (B)	.47	2	.235	13.106*
Trial (T)	1.22	7	.174	4.539*
Error S(G)	1.153	6	.192	
G x B	.054	2	.027	1.518
G x T	.203	7	.029	.756
B x T	3.133	14	.224	7.04*
Error SB(G)	.215	12	.018	
Error ST(G)	1.613	42	.038	
G x B x T	.351	14	.025	.788
Error SBT(G)	2.67	84	.032	

* $p < .005$

agreement; the careful group, .58. Because high accuracy by both members of an observer pair means that agreement is more likely to also be high than if only one member of the pair achieves high accuracy, these findings may be interpretable only in the light of the findings regarding accuracy.

CHAPTER IV

DISCUSSION

Summary of Major Findings

The present study attempted to answer the following questions about the effect of certain variables on the behavior of human observers: first, what effect, if any, do instructions to observers have on the levels of agreement and accuracy they achieve; and, second, do instructions and whether the data are their own or those of another observer pair produce differences in the way observers calculate inter-observer agreement.

The findings with regard to the first question were that instructions did indeed affect the agreement/accuracy relationship, though they did not affect either of these measures separately. When instructions emphasized the importance of high inter-observer agreement, inter-observer agreement exceeded observer accuracy; when the instructions emphasized careful, independent recording of behavior, the relationship was reversed. Group differences due to instructions were significant.

With regard to the second question, the findings were that all observers, regardless of the instructions they were given, spuriously inflated their own agreement levels and spuriously deflated the agreement levels of data they

thought were collected by other observers. These differences were significant.

Errors in Calculating Inter-Observer Agreement

Kent and Foster (1977) note that the implicit demand for observers to produce high levels of inter-observer agreement may result in their "cheating" in order to meet this demand. Such "cheating" may occur if observers communicate during the collection of behavioral observations, modify their records during assessment of agreement, or make calculation errors which spuriously inflate reported agreement levels.

In the present study, the first two types of "cheating" were presumably controlled. The experimenter remained in the laboratory at all times and asked observers not to talk to each other. Observers' data were duplicated at the time the data were collected by using carbon paper. Since one copy of the data were exchanged, observers could not have modified both records.

Agreement calculations, however, were not closely monitored, in order to study the effects of instructions and the experimental procedure on calculation errors and the subsequent agreement levels reported by observers. Previous research on calculation errors manipulated several different independent variables and produced somewhat inconsistent results. One independent variable is whether the observers or the experimenter calculated agreement scores. On one hand, Rusch et al. (1975) reported that the calculation errors made by two

experimental assistants summarizing the data collected by others were small, were non-directional (not consistently in the direction of the experimental hypothesis), and inconsequential with respect to the reported treatment outcome. On the other hand, O'Leary and Kent (1973) found that observers calculating agreement levels for their own data reported agreement levels which, when re-calculated by the experimenters, were spuriously inflated by eight points (agreement levels of .66 and .58, respectively). Kent et al. (1974) found six point differences in subject-calculated and experimenter-calculated agreement levels for two groups of observers calculating agreement for their own data.

Subject versus experimenter calculation of agreement levels thus has been generally shown to affect the levels of agreement reported for observational data. The presence or absence of the experimenter from the room during subject calculation of agreement has also been shown to have an effect on reported agreement levels. O'Leary and Kent (1973) found that average observer agreement of .66 on days when the experimenter was absent dropped to .55 when he was present. Kent, Kanowitz, O'Leary, and Cheiken (1977) found that experimenter absence from the room during agreement assessment inflated agreement scores by an average of six points over scores obtained in the experimenter's presence.

The present study is noteworthy because it demonstrated that another variable affected reported agreement levels in

addition to the two mentioned above. In addition to experimenter presence or absence and subject versus experimenter calculated agreement levels, the agreement levels reported by observers were affected by whether the calculations were done by the subject on his or her own data or on the data of a different pair of observers. If the data were the subject's own, the agreement levels were spuriously inflated; if the data were those of other observers, the agreement levels were spuriously deflated.

One explanation for these findings requires that one examine the antecedents and consequences of subjects' behavior in calculating levels of inter-observer agreement. The analysis assumes that the experimental setting itself is somewhat anxiety-producing, and that this anxiety is heightened when subjects evaluate their own performance (calculate inter-observer agreement) and discover that other subjects are performing at higher levels than they are. In the present study the agreement levels pre-established for the contrived ("other" observers') data were higher than the levels most observers in the study were able to achieve. One might hypothesize that a reduction in anxiety might be reinforcing to observers. One way that observers could reduce this anxiety would be to reduce differences in their own level of performance and that of "other" observers. They could do this either by spuriously inflating their own levels of agreement, or by spuriously deflating the levels of agreement achieved by "others." In the present study, observers did both.

The Relationship Between Observer Accuracy and Inter-Observer Agreement

When standards which reflect the "true" occurrence of behavior are not available (which is the case in most natural and even contrived settings), observers' records are usually compared to each other, and levels of inter-observer agreement are calculated. High levels of inter-observer agreement are used as the best evidence that the observations probably reflect the occurrence of behavior.

However, the relationship between observer accuracy and inter-observer agreement appears to be complex. Limited knowledge about this relationship makes it impossible to predict accuracy scores by knowing the agreement scores. In addition, little is known about the variables which affect the relationship between observer accuracy and inter-observer agreement.

In the present study, for example, instructions to observers had a significant effect on this relationship. When the instructions emphasized that agreement levels should be as high as possible, agreement scores were higher than accuracy scores (by .05). When instructions emphasized careful observation and made no mention of the need for high agreement, accuracy scores were found to be higher than agreement scores (by .02). These differences were significant ($p < .05$).

Kapust and Nelson (Note 7) studied the relationship between observer accuracy and inter-observer agreement in a vigilance analogue to naturalistic observation. Subjects in

this study counted two arbitrary behaviors (lifting and/or moving the index finger of the hands of experimental assistants); the rates of occurrence of each response were preprogrammed in order to establish a standard against which observer accuracy could be assessed. The results indicated that observer accuracy and agreement are not linearly related; further, values of the two variables were often discrepant by more than ten points (and by 26 points at the extreme).

The implications of the present study and that of Kapust and Nelson are that researchers should not necessarily assume that certain levels of inter-observer agreement imply certain levels of accuracy. Rather, an attempt should be made to assess the accuracy of observers' records. Unobtrusive videotaping of a portion of an experiment might provide a standard against which observers' records could be compared. The use of electronic measurement might also provide a standard, as Kapust and Nelson pointed out.

Instructions to Observers

Several variables have been shown to affect how closely an observer's data record compares to that of another observer (inter-observer agreement) or to a standard which supposedly more nearly reflects the actual occurrence of the behavior (observer accuracy): complexity of the behavior observed (Jones et al., 1974); code complexity (Mash & McElwee, 1974); knowledge of assessment of agreement (Reid, 1970; Taplin & Reid, 1973); familiarity with the assessor of agreement

(Romanczyk et al., 1973); the effects on coding new behavioral sequences following coding predictable versus unpredictable behavioral sequences during training (Mash & McElwee, 1974); and experimenter status (Taplin & Reid, 1973).

The present study attempted to demonstrate the effects on agreement and accuracy of a new variable, instructions. One half of the subjects were instructed to try to achieve high agreement with their partner. This group was expected to achieve a significantly higher overall level of agreement, but a significantly lower overall level of accuracy, than the group instructed to observe and record their data carefully.

Although the directions of group differences on these two measures were as predicted, in neither case were the differences significant at conventional significance levels (group differences in accuracy were significant at the .10 alpha level). The particular experimental procedures may have minimized in two ways the effects of instructions on the two groups. First, because the experimenter was always in the room, and also asked subjects not to talk to each other, subjects in the high agreement demand group may not have been able to determine how their partner was applying the behavioral code.

The second way that the effects of instructions may have been minimized relates to the training procedure that was used. Johnson and Bolstad (1973) note that training may provide observers with the opportunity to learn how their

partner is applying the code, especially if the two observers are trained as a team. The result may be an idiosynchratic application of the code by the observer pair, called "consensual observer drift," and measured either by comparing agreement levels achieved within an observer pair to those obtained between different pairs, or by comparing an observer pair's agreement levels to their accuracy levels. In the present study, an attempt was made to minimize the possibility of "drift" during training. Observers were trained as a single group, and observers who tended to be present during the same training sessions were not paired.

The Pattern of Observer Accuracy Across Trials

Observers completed four observational trials on each of two experimental days. The statistical analysis of observers' accuracy levels revealed a significant main effect for trials. The pattern of the results indicated that observers' accuracy for each experimental day showed a decline across the first three trials, then an increase in accuracy on the fourth (final) trial of the day. Observational accuracy was highest on the final trial of each day; the Scheffé post-hoc comparison among means detected the significant difference in accuracy between the last trial of the day and the trial which immediately preceded it.

The pattern of results across the first three trials of each day suggests that observers may have become fatigued or

bored as the session progressed. But why, then, the increase to the highest level of observational accuracy on the last trial of the day?

One possible explanation for these findings utilizes Hilgard and Bower's (1966) description of the "gradient of reinforcement." It is first assumed that the behavioral requirements made of observers on an experimental day constitute a behavioral chain. It is also assumed that the reinforcer for the observers in this setting is the opportunity to leave the laboratory. The gradient of reinforcement explanation suggests that an organism's performance is enhanced just prior to reinforcement. This could explain the enhancement of observational accuracy just before the end of the session.

One possible methodological confound exists which may temper the findings of differences in accuracy across trials. Eight different segments of videotape were presented to the subjects across the eight trials; however, all the subjects viewed the tapes in the same sequence. This means that tape content, not trial, might account for the differences in observational accuracy across trials.

Conclusions

Wildman and Erickson (1977) have noted and documented the historical reliance of psychological research on human observers collecting behavioral data in natural settings.

The increasing popularity of behavioral assessment and intervention strategies during the 1960's and 1970's has led to an even greater reliance on data collected by human observers. Thus, it is not surprising that many researchers are now concerned with studying the variables which affect the data collected in this manner.

Methodological research in this area has typically addressed itself to four major issues: changes in the behavior of subjects as a result of being observed (reactivity); characteristics of the instrument (usually some type of behavioral coding system) used by observers; procedures related to assessing the reliaiility of human observation (inter-observer agreement); and observer biases in collecting data.

This study has contributed to the overall findings regarding inter-observer agreement and how it relates to observational accuracy. Thus, it is a study of the validity of research which uses data collected by human observers. Not only is it crucial that the levels of agreement reported by observers be correctly calculated, it is also important that the levels of agreement achieved by observers reflect a level of observational accuracy which will give meaning to the results of the study, both in terms of the relation between independent and dependent variables and the generalizability of the findings.

This study demonstrated that observers calculating inter-observer agreement made calculation errors which spuriously inflated reported agreement levels (if the data were the observer's own), or spuriously deflated them (if the data were those of other observers). These findings appear to support Kazdin's (1977) precautions that agreement levels be calculated by persons other than those who collect the data.

The study also demonstrated that instructions to observers affected the relationship between the observers' agreement with their partner and their observational accuracy. The agreement/accuracy relationship has provided one means for studying "consensual observer drift." Consensual observer drift occurs when a pair of observers changes, over time, the way they apply a behavioral code. Drift is often measured by comparing the agreement levels within pairs of observers to those between different pairs of observers. When standards reflecting the actual occurrence of behavior are available, drift may be examined by comparing levels of inter-observer agreement to individual observer accuracy. This is because observer drift allows observers working as a pair to maintain acceptable levels of agreement, despite a decline in the accuracy of their observations. In the present study, instructions creating a demand for high agreement resulted in observer agreement exceeding observer accuracy;

when the instructions created a demand for careful observation, the relationship was reversed. These findings seem to have implications regarding how observers are trained, especially the extent to which the importance of inter-observer agreement is emphasized.

Summary

In addition to making observational recordings of behavior, observers in this study calculated their own levels of inter-observer agreement; they also calculated agreement levels on data they were led to believe were collected by other observers in the study. This procedure resulted in observers reporting agreement levels for their own data which, when re-calculated by the experimenter, were found to be spuriously high. The agreement levels they reported for the "other" observers' data were found to be spuriously low.

These findings suggest that, regardless of the instructions observers are given, they perceive the calculation of agreement to be an evaluation of their performance. Their behavior in this study had the effect of making their performance appear to be as nearly like that of "other" observers as possible.

The implications of these findings for behavioral research suggest that researchers not have their observers calculate agreement levels; rather, assistants not otherwise involved in the study might be asked to perform the necessary calculations.

The study examined inter-observer agreement and its relationship to individual observer accuracy. Group differences resulting from different instructions were found to have an effect on the relationship between these two measures. When instructions emphasized the need for high agreement, overall agreement levels exceeded levels of accuracy. When instructions emphasized careful and independent observational recording, the relationship was reversed.

These findings seem to have implications for training of observers. Over-emphasizing the importance of high inter-observer agreement may result in observers sacrificing individual accuracy in order to achieve and maintain high agreement. The end result may be data which are reliable but which are inaccurate to the point of invalidating the results of the experiment. The findings seem to suggest that criteria for assessing observational accuracy be developed whenever possible, perhaps by videotaping a sample of the observational sessions during the experiment.

Further research is certainly indicated to gain a better understanding of the variables which affect the relationship between agreement and accuracy. This is especially true since in most studies that utilize observational data, agreement levels are assumed to be the best index of the accuracy of the data, and thus the validity of the results.

REFERENCE NOTES

1. Kass, R., & O'Leary, K. D. The effects of observer bias in field-experimental settings. Paper presented at a symposium, Behavior analysis in education, University of Kansas, Lawrence, Kansas, April 9, 1970.

2. O'Leary, K. D., Romanczyk, R. G., Kass, R. E., Dietz, A., & Santogrossi, D. Procedures for classroom observation of teachers and children. Unpublished manuscript, Point-of-Woods Laboratory School, State University of New York at Stony Brook, 1971.

3. Skindrud, K. An evaluation of observer bias in experimental-field studies of social interaction. Unpublished doctoral dissertation, University of Oregon, Eugene, 1972.

4. Kent, R. The human observer: An imperfect cumulative recorder. Paper presented at the Fourth Banff Conference on Behavior Modification, Banff, Alberta, Canada, 1972.

5. Fogel, L. Personal communication, October, 1976.

6. Kent, R. N., O'Leary, K. D. Coletti, G., & Drabman, R. S. Increasing the validity of an observational code: An empirical methodology. Manuscript in preparation, 1975.

7. Kapust, J. A., & Nelson, R. O. Parameters influencing inter-observer agreement and observer accuracy in a vigilance analogue to naturalistic observation. Manuscript in preparation, 1976.

BIBLIOGRAPHY

Azrin, N., Holz, W., Ulrich, R., & Goldiamond, I. The control of conversation through reinforcement. Journal of the Experimental Analysis of Behavior, 1961, 4, 25-30.

Barber, T. X. Pitfalls in human research: Ten pivotal points. New York: Pergamon Press, Inc., 1976.

Barber, T. X., & Silver, M. J. Fact, fiction, and the experimenter bias effect. Psychological Bulletin (Monograph Supplement), 1968, 70 (No. 6, Pt.22), 1-29.

Bechtel, R. B. The study of man: Human movement and architecture. Transaction, 1967, 4, 53-56.

DeMaster, B., Reid, J., & Twentyman, C. The effects of different amounts of feedback on observers' reliability. Behavior Therapy, 1977, 8, 317-329.

Hilgard, E. R., & Bower, G. H. Theories of learning. New York: Appleton-Century-Crofts, 1966.

Jastak, J. F., Bijou, S. W., & Jastak, S. R. Wide range achievement test: Reading, spelling, arithmetic from pre-school to college. Wilmington, Delaware: Guidance Associates of Delaware, Inc., 1965.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts, and practice. Champaign, Ill.: Research Press, 1973.

Jones, R. R., Reid, J. B., & Patterson, G. R. Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 3). San Francisco: Jossey-Bass, 1974.

Kazdin, A. E. Artifact, bias, and complexity of assessment: The abc's of reliability. Journal of Applied Behavior Analysis, 1977, 10, 141-150.

Kent, R. N., & Foster, S. L. Direct observational procedures: Methodological issues in natural settings. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), Handbook of behavioral assessment. New York: John Wiley & Sons, 1977.

Kent, R. N., Kanowitz, J., O'Leary, K. D., & Cheiken, M. Observer reliability as a function of circumstances of assessment. Journal of Applied Behavior Analysis, 1977, 10, 317-324.

Kent, R. N., O'Leary, K. D., Diament, C., & Dietz, A. Expectation biases in observational evaluation of therapeutic change. Journal of Consulting and Clinical Psychology, 1974, 42, 774-780.

Lipinski, D., & Nelson, R. Problems in the use of naturalistic observation as a means of behavioral assessment. Behavior Therapy, 1974, 5, 341-351.

Mash, E. J., & McElwee, J. D. Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. Child Development, 1974, 45, 367-377.

Mischel, W. Personality and assessment. New York: John Wiley and Sons, Inc., 1968.

O'Leary, K. D., & Kent, R. N. Behavior modification for social action: Research action: Research tactics and problems. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts, and practice. Champaign, Ill.: Research Press, Inc., 1973.

O'Leary, K. D., Kent, R. N., & Kanowitz, J. Shaping data congruent with experimental hypotheses. Journal of Applied Behavior Analysis, 1975, 8, 43-51.

Pawlicki, R. Behaviour therapy research with children: A critical review. Canadian Journal of Behavioral Science, 1970, 2, 163-173.

Reid, J. B. Reliability assessment of observation data: A possible methodological problem. Child Development, 1970, 41, 1143-1150.

Repp, A. C., Dietz, D. E. D., Boles, S. M., Dietz, S. M., & Repp, C. F. Differences among common methods for calculating inter-observer agreement. Journal of Applied Behavior Analysis, 1976, 9, 109-113.

Roberts, R. R., & Renzaglia, G. A. The influence of tape recording in counseling. Journal of Counseling Psychology, 1965, 12, 10-16.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. Measuring the reliability of observational data: A reactive process. Journal of Applied Behavior Analysis, 1973, 6, 175-184.

Rosenthal, R. On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. American Scientist, 1963, 51, 268-283.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, Inc., 1966.

Rosenthal, R., & Fode, K. L. The effect of experimenter bias on the performance of the albino rat. Behavior Science, 1963, 8, 183-189.

Rosenthal, R., & Jacobsen, L. Teacher's expectancies: Determinants of pupils' IQ gains. Psychological Reports, 1966, 19, 115-118.

Rusch, F. R., Walker, H. M., & Greenwood, C. R. Experimenter calculation errors: A potential factor affecting interpretation of results. Journal of Applied Behavior Analysis, 1975, 8, 460.

Scott, P., Burton, R. V., & Radke-Yarrow, M. Social reinforcement under natural conditions. Child Development, 1967, 38, 53-63.

Shuller, D. Y., & McNamara, J. R. Expectancy factors in behavioral observation. Behavior Therapy, 1976, 7, 519-527.

Taplin, P. S., & Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. Child Development, 1973, 44, 547-554.

Thomson, C., Holmberg, M., & Baer, D. M. A brief report on a comparison of time-sampling procedures. Journal of Applied Behavior Analysis, 1974, 7, 623-626.

Weber, S. J., & Cook, T. D. Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. Psychological Bulletin, 1972, 77, 273-295.

Wildman, B. G., & Erickson, M. G. Methodological problems in behavioral observation. In J. D. Cone & R. P. Hawkins (Eds.), Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel, Inc., 1977.

Wildman, B. G., Erickson, M. T., & Kent, R. N. The effect of two training procedures on observer agreement and variability of behavior ratings. Child Development, 1975, 46, 520-524.

APPENDIX A

Arithmetic Screening Test

(Items from Wide Range Achievement Test)

1. $\begin{array}{r} 43 \\ +\ 6 \\ \hline \end{array}$

2. $\begin{array}{r} 94 \\ -64 \\ \hline \end{array}$

3. $\begin{array}{r} 726 \\ -349 \\ \hline \end{array}$

4. 1/6 of 30 = _____

5. $\begin{array}{r} 229 \\ 5048 \\ 63 \\ +1381 \\ \hline \end{array}$

6. $9\overline{)4527}$

7. Write as percent:

   .42 = _____%

8. Find average:

   34, 16, 45, 39, 27

   Ans._____

9. Write as decimal:

   $52\frac{1}{2}\%$ = _____

10. Write as percent:

    3/8 = _____%

APPENDIX B

Order of Presentation of Data Sets

Observer pairs	Trial 1	2	3	4	5	6	7	8
1	A	B	B	A	A	B	B	A
2	B	A	A	B	B	A	A	B
3	B	A	A	B	B	A	A	B
4	A	B	B	A	A	B	B	A
5	B	A	A	B	B	A	A	B
6	A	B	B	A	A	B	B	A
7	A	B	B	A	A	B	B	A
8	B	A	A	B	B	A	A	B

A - observer pair receives "own" data first, then "other" data

B - observer pair receives "other" data first, then "own" data

(odd numbered observer pairs are in the same instructions group; even numbered pairs are in the same instructions group)

APPENDIX C

OBSERVER'S INSTRUCTIONS AND TRAINING PACKAGE

I. Purpose of the study

This is a study of the effects of a behavioral treatment program on children's classroom behavior. The teacher selected two children with whom she has had difficulty in working. We videotaped the children on several occasions before and after the treatment program was implemented. Your job will be to observe some of the videotapes, record your observations of the children's behavior, and perform some basic arithmetic computations on the data collected by you and by other observers.

The use of videotapes should prevent your learning two facts, both of which would be obvious to you if you observed in the classroom: first, which of the two children was receiving the treatment program at any given time; and second, whether the treatment program had begun or not. Thus, we can eliminate the possible effects of your knowing these facts on your observations by using videotapes. You are encouraged to use the "Comments" section of the data sheets to relate your speculations regarding these or any other aspects of the study.

II. Training procedures

Before you begin observing the experimental tapes, you will be trained to use the behavior coding system and to perform the proper arithmetic calculations. Two training

sessions will be scheduled; you will be given a copy of the schedule. Pleaee notify the experimenter as soon as possible if you become unable to attend either of these sessions.

You will receive detailed operational definitions of the target responses you will code from the tapes. Ask the experimenter to clarify any confusing areas of the definitions. You will also observe a six-minute tape selected for training purposes. The children whose behavior you will later record will be shown emitting each of the responses you are to record. The experimenter will stop the tape periodically to describe what you have just seen. Also, the experimenter will provide a data sheet on which are recorded hypothetical observations of the behaviors for that segment of the tape.

The experimenter will then describe the observational procedure (called time sampling) you will use throughout the study. Basically, this is the procedure: the tape is divided into a number of small time intervals; observers spend a portion of each time interval observing the tape, and the remainder of the interval recording their observations. In this study, the experimental tapes are ten minutes long. Each tape is divided into twenty intervals of equal length (in this case, thirty seconds). You will spend the first twenty seconds observing the tape, and the last ten seconds recording your observations. A tape recording will tell you which time interval you are working on, when you should observe, and when you should record your observations.

After you have viewed the demonstration tape and heard an explanation of the observational recording procedure, you will be asked to refer to a second data sheet in your instructions package. Hypothetical data are recorded on it; this data sheet has been compared to the other data sheet in your package in order to show you how to calculate inter-observer agreement between two data sheets. Look at the portion of the data sheet where inter-observer agreement figures are calculated. The formula used for calculating agreement is: Agreement intervals ÷ Agreement intervals plus Disagreement intervals. An agreement interval is counted whenever both data sheets indicate that the behavior occurred during that interval. A disagreement interval is counted whenever only one data sheet indicates that the behavior occurred. Intervals where neither data sheet indicate the occurrence of the behavior are not counted in the calculations. This means the denominator in the formula will not necessarily add up to the number of recording intervals.

Next you will get a chance to practice the coding procedure. You will receive blank data sheets and will record your observations of the behavior of one of the children from a six-minute training tape. The experimenter will also collect data from this tape, and will answer any questions at the end of it.

After this, you will practice the procedure for calculating inter-observer agreement. You will note that the data sheets are equipped with carbon paper so that a copy of your

data are made. This will allow you to exchange your data with someone else so that everyone will have two data sheets to compare. The experimenter will answer any questions you have about this procedure.

At the second training session you will continue to practice using the code to record your observations from training tapes. You will also continue to practice calculating inter-observer agreement.

III. Experimental procedure

You will be scheduled to come to the lab four times for thirty minutes each time. During the thirty minutes you will observe two ten-minute videotaped segments of the children's classroom behavior. You will collect data on one of the children from the tape. Then you will calculate agreement figures for that session, both on the data you and your partner collect, and on data collected by other pairs of observers during the study. You will use the same observation, recording, and calculation procedures used during training.

Classroom Intervention Study

Vocalization--symbol = V

Purpose: Vocalization is intended to monitor verbal behavior which is usually distracting to both the child and to others.

Description: For the sake of consistency, any audible nonpermitted vocalization is to be recorded even though in the opinion of the observer it did not "seem" disruptive. Any non-permitted "a ible" behavior emanating from the mouth.

Critical Points: The observer must actually hear the vocalization. Inferences are not acceptable except as noted below.

If vocalization is obvious, but can't be heard (obvious--if another child responds). Answering without being called on. Moaning. Yawning. Any noise made with nouth when eating--unless child has permission to eat. Any vocalization made in response to the disruptive behavior of another child, e.g., telling another child to return stolen article, crying in response to aggression committed to his person or possessions, etc., if the child has not received permission specifically from the teacher to speak.

Whispering, Belching, Crying, Shouting, "Operant" coughs or sneezes.

Excludes: Vocalization in responses to teacher's question. Sneezing. Automatic coughing.

Note: Once a child is recognized by the teacher, vocalization is not scored, regardless of content of the vocalization: crying, yelling, swearing, etc., until the teacher specifically instructs the child to stop.

Playing--symbol = P

Purpose: Playing is intended to monitor often subtle manipulative behavior that is distracting to the child and possibly also distracting to others.

Description: Child uses his hands to play with his own or community property, so that such behavior is incompatible (or would be incompatible) with learning.

Critical Points: Child uses his hands to manipulate his own or community property.

Includes: Playing with toy car when assignment is spelling. Playing with comb or pocket book. Eating only when the hands are being used--chewing gum is not rated as P unless child touches or manipulates it with his hands. Poking holes in workbook. Cleaning nails with pencil. Drawing on self. Manipulating pencil in such a manner as to make the behavior incom-

patible with learning, e.g., shoving pencil back and forth on desk; waving pencil through air as an airplane. Picking scabs, nails, or nose if the desired "object" is separated from the body and manipulated. Looking into desk and moving arms, but does not come out with a task-related object. Working with or reading non-task related material, e.g., reading page 25 when told to read page 1, doing math when told to do spelling, etc.

Excludes: Touching others' property. Playing with own clothes.

Note: Include if article is removed from body, e.g., shoes, tie, buttons, scarf, etc., and is manipulated.

Lifting desk or chair with feet (rate N if this creates audible noise). Random banging of pencil on desk (rate N if audible). Simple twiddling pencil if it is not seen as being incompatible with learning.

Note: Rate twiddling pencil, banging pencil, or putting pencil in mouth, hair, behind ear, etc., if child attends to such behavior and ceases attending to assigned task. Operational definition of attending: child either looks at manipulated object or begins to manipulate object in non-random patterns for more than 5 seconds.

Picking scabs, nails, or nose if the desired "object" is not separate from the body.

Orienting Response--symbol =

Purpose: Orienting is intended to monitor the gross motor behavior of turning around from the designated point of reference. Such behavior is distracting to child since it usually precludes attending to assigned task, and is often distracting to others.

Description: Child turning more than 90 degrees from point of reference while seated.

Critical Points: The child must be in his seat; he may be in a modified position; and orienting includes both the horizontal and vertical axis.

Includes: Turning to the person behind. Looking to the rear of the room. Turning around in chair or turning chair around. Leaning back in chair more than 90 degrees.

Note: Point of reference is typically child's desk, but may be the teacher if the children are directed to attend to her. If child should turn desk at some angle, point of reference becomes where desk was originally, not to where the child has moved it. Also, the child's chin should be used as the indicator of how far he has turned.

Therefore, orienting is noted when child's chin has turned more than 90 degrees from point of reference.

Excludes: Orienting during class discussion when the teacher directs (either implicitly or explicitly) the class to attend to a child's explication of an answer. Orienting while picking up a task related object. When child is in corner or otherwise out of his chair.

Classroom Intervention Study

Data Sheet

Observer's Initials ABC Videotape Number 712

Today's Date 7/12/77 Session Number P

Time 1:00 PM Initials of Other Observer XYZ

****

Observational Data

Interval

1	2	3	4	5	6	7	8	9	10
~~P~~	~~P~~	~~P~~	~~P~~	~~P~~	~~P~~	~~P~~	~~P~~	~~P~~	~~P~~
V	V	V	~~V~~	~~V~~	V	V	V	~~V~~	~~V~~
~~Ø~~	~~Ø~~	O	~~Ø~~	O	~~Ø~~	O	O	~~Ø~~	O
P	P	P	P	P	P	P	P	P	P
V	V	V	V	V	V	V	V	V	V
O	O	O	O	O	O	O	O	O	O
11	12	13	14	15	16	17	18	19	20

Interval

****

Data Summary

Frequency Inter-observer Agreement Calculations Agreement

Play 10	Play	Play
Vocal 4	Vocal	Vocal
Orient 5	Orient	Orient

****

Comments

Classroom Intervention Study

Data Sheet

Observer's Initials XYZ Videotape Number 712

Today's Date 7/12/77 Session Number P

Time 1:00 PM Initials of Other Observer ABC

****

Observational Data

Interval

1	2	3	4	5	6	7	8	9	10
~~P~~	~~P~~	~~P~~	P	~~P~~	~~P~~	P	~~P~~	~~P~~	~~P~~
~~V~~	V	V	~~V~~	~~V~~	V	~~V~~	V	~~V~~	~~V~~
~~O~~	O	O	~~O~~	O	~~O~~	O	O	~~O~~	O
P	P	P	P	P	P	P	P	P	P
V	V	V	V	V	V	V	V	V	V
O	O	O	O	O	O	O	O	O	O
11	12	13	14	15	16	17	18	19	20

Interval

****

Data Summary

Frequency	Inter-observer Agreement Calculations	Agreement
Play 8	Play: $\frac{\text{Agree}}{\text{Agree + Disagree}} = \frac{8 \text{ (Intervals 1,2,3,5,6,8,9,10)}}{8 + 2 \text{ (Intervals 4,7)}}$	Play .80
Vocal 6	Vocal: $\frac{4 \text{ (Intervals 4,5,9,10)}}{4 + 2 \text{ (Intervals 1,7)}} = \frac{4}{6} = \frac{2}{3} = .67$	Vocal .67
Orient 4	Orient: $\frac{4}{4 + 1} = \frac{4}{5} = .80$	Orient .80

****

Comments

APPENDIX D

Classroom Intervention Study

Questionnaire

(please circle appropriate letter)

1. Which of the three target behaviors do you feel was the easiest for you to observe and record?

   a. Playing
   b. Vocalizing
   c. Orienting
   d. all about equal

2. What (if any) instructions were you given regarding your own observations at the beginning and at the halfway point each day?

   a. no instructions given
   b. try to reach .85 inter-observer agreement levels with partner, if possible
   c. act as an independent recording device by making my observations and calculations as carefully as possible
   d. b and c

   e. neither b nor c, but some other instructions (please specify)

   _____

   _____

   f. don't remember instructions

3. How accurate do you feel your observations were, in general, over time?

   a. my observations became probably more accurate with each session
   b. my observations became probably less accurate with each session
   c. my observations were probably equally accurate over all sessions
   d. no pattern
   e. I don't know

APPENDIX E

Instructions to Subjects in High Agreement Group

(to be read after experimenter answers any procedural questions, then asks subjects not to speak to each other until the final trial of that day is over)

Today you will observe four 10-minute videotaped segments of the classroom behavior of two second-grade children and record the occurrence of the target behaviors we are interested in studying. At the conclusion of each segment I will collect your data sheets. Then I will give each of you two data sheets and ask you to calculate inter-observer agreement for each target behavior. You may do the calculations on either data sheet. When you have finished I will collect the data sheets, give you a second set of data, and ask you to again calculate agreement figures for each behavior. One set of data will be your own; the other will be those data collected by other observers in the study.

Let me emphasize one thing about your own observations. One way researchers judge the quality of the data collected by observers is by how closely the records of the observers compare. Minimal inter-observer agreement levels of .85 indicate that the data collected are of sufficient quality to be used in reporting the results of a treatment program. Hopefully, you and your partner will achieve a .85 level of agreement on most of your observations.

(experimenter tells subjects which child to observe, then instructs them to begin observing when cued by the tape)

Instructions to Subjects in Careful Group

(to be read after experimenter answers any procedural questions, then asks subjects not to speak to each other until the final trial of that day is over)

Today you will observe four 10-minute videotaped segments of the classroom behavior of two second-grade children and record the occurrence of the target behaviors we are interested in studying. At the conclusion of each segment I will collect your data sheets. Then I will give each of you two data sheets and ask you to calculate inter-observer agreement for each target behavior. You may do the calculations on either data sheet. When you have finished I will collect the data sheets, give you a second set of data, and ask you to again calculate agreement figures for each behavior. One set of data will be your own; the other will be those data collected by other observers in the study.

Let me emphasize one thing about your own observations. One way researchers judge the quality of the data collected by observers is by how carefully the experimenter feels each observer is collecting his or her own data. Observers are usually asked to consider themselves as independent recording devices in order for the data they collect to be of high quality. The frequency data you and your partner collect are of primary interest to me. These data are averaged. So please try to perform your observational recordings and calculations as carefully as possible.

(experimenter tells subjects which child to observe, then instructs them to begin observing when cued by the tape)

APPENDIX F

Agreement Levels for Hypothetical Data Sets

(one data sheet will be derived from the protocol for that videotaped segment; the other will be drawn up so that the following levels of inter-observer agreement will be obtained if calculations are performed correctly):

Target Behavior

Experimental Session	Playing	Vocalizing	Orienting
#1	.92	.86	.70
#2	.75	.83	.86
#3	.81	.83	.92
#4	.94	.90	.86
#5	.83	.80	.77
#6	.88	.80	.80
#7	.92	.83	.67
#8	.73	.80	.86